JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1998, **31,** 263–280 NUMBER 2 (SUMMER 1998)

WHEN PERFORMANCE IMPROVEMENT IS THE GOAL: A NEW SET OF CRITERIA FOR CRITERIA

JUDITH L. KOMAKI

BARUCH COLLEGE, CITY UNIVERSITY OF NEW YORK

Weitz (1961), aware of the lure of tradition and expedience, urged the identification of evaluation standards for dependent variables, which he referred to as criteria for criteria. In this article, five criteria are proposed using the mnemonic SURF & C: the directness of *sampling* (S); the responsiveness of the target (i.e., making sure the dependent variable is *under* (U) the worker's control); the *reliability* (R) of observers; the *frequent* (F) assessment of the target during the intervention period; and the *critical* (C) nature of the target. Together the criteria provide guidelines for what and how targets should be assessed. Their necessity is illustrated in two year-long experiments designed to improve the preventive maintenance of heavy equipment in the U.S. Marine Corps. Although the criteria are limited to evaluating dependent variables in field experiments, they are recommended as the foundation for successful performance efforts in any applied setting.

DESCRIPTORS: performance appraisal, observation, measurement, assessment, dependent variable, target behavior, feedback

In an article entitled "Criteria for Criteria," Weitz (1961) posed the tantalizing question of whether our results would be different if we used one operational definition rather than another.

> If . . . we are evaluating the effectiveness of two simulators as training devices, would our conclusions concerning the utility of these simulators vary depending upon whether we chose as our criterion: the length of time required on each simulator to reach some level of performance on the actual instrument itself, the number of errors on the actual device after N trials on each simulator, the speed with which a certain level of performance on the actual device can be reached, or the performance of individuals on the actual device 6 months after simulator training? (p. 228)

Weitz argued that the choice of the operational definition makes a crucial difference. To make the point, he presented data from a learning study showing how changing the target altered the findings and the eventual interpretation. He concluded by calling for standards for dependent variables, which he referred to as *criteria for criteria.*

Just as we have criteria for selecting restaurants (e.g., a rating above 20 for food quality; see *Zagat Survey,* 1996) or for identifying estimators in classical statistical theory (e.g., lack of bias, small variance, linearity), so we can have criteria for selecting targets. The term *target* refers to the dependent variable and its operational definition (Underwood, 1957) or the latent variable and its manifest measure (Borman, 1991). Hence, it can refer to training effectiveness

This research was partially supported by U.S. Office of Naval Research and Naval Personnel and Research and Development Contracts N00014-79-C-0011 and NR170-9918-25-78. The views, opinions, and findings contained in this paper are those of the author and should not be construed as an official Department of the Navy position, policy, or decision. I thank the technical monitors, Bert King, Robert Hayles, and Al Lau; colleagues, students, and friends, Milton Blood, Jeff Daum, Anthony DeCurtis, Larry James, Rich Kopelman, Jennifer Scharer, Pat Smith, Tom Redding, and Beth Sulzer-Azaroff; observers, Alex Chacto, Swede Larson, and Don Hermann; and the Marines of the Second Marine Division.

Address correspondence to the author at the Department of Psychology, Baruch College, CUNY, 17 Lexington Avenue, Box G1126, New York, New York 10010.

or operational definitions such as the time to complete the task, the number of errors, the speed with which trainees attain a certain level of performance during training, or their performance on the job. *Target* is not restricted to behaviors, but can include the outcomes of these behaviors as well. Hence, making errors as well as the errors themselves could be considered to be targets. (Ironically, the areas of criterion measurement and construct validity are plagued with poorly defined constructs. The term *performance,* for example, is sometimes equated with outcomes; in other contexts, it refers to both behaviors and outcomes. Paralleling the custom in applied behavior analysis, I use the term *target* to refer to both.)

In this article, I review the literature in the fields of applied behavior analysis (ABA) and industrial/organizational (I/O) psychology, pointing out problems with many appraisals of performance. I then propose five criteria for criteria. To illustrate the necessity of judiciously identifying both what and how targets should be assessed, I describe what I learned from developing targets in a case study. The last section addresses the contribution of the criteria to the area of criterion measurement and ends with recommendations for future research.

PERSISTENT CRITICISMS

Despite Weitz's (1961) call more than 35 years ago, researchers in the field of I/O psychology still complain about the quality of performance appraisals in work settings. In contrast to tests that Bernardin and Beatty (1984) characterized as "careful observations of actual performance under standard conditions," performance appraisals were depicted as "careless observations of performance under unstandardized conditions" (p. 128). In their lengthy review, Austin, Villanova, Kane, and Bernardin (1991) portrayed the construct validation of performance measures as a "flat learning curve" (p. 215).

Dissatisfaction was also expressed by researchers in applied behavior analysis (Finney, 1991; Sulzer-Azaroff & Fellner, 1984). In their memorably titled article, "Be Still, Be Quiet, Be Docile," Winett and Winkler (1972) chided professionals for neglecting targets that are more educationally relevant. Weist, Ollendick, and Finney (1991) pointed to such dubious practices as basing the behaviors on convenience or standard practice, choosing behaviors along erroneous or irrelevant dimensions, and limiting assessments to self-reports and interviews. In a comprehensive chapter on direct observation, Foster and Cone (1986) raised questions about the observers: Are observers' definitions of targets likely to drift and gradually change over time? Will their expectations about who and what they are observing bias their results?

These criticisms are not limited to academics. Employees complain as well. Some have successfully sought compensation in court (e.g., *Price Waterhouse v. Hopkins,* 1989), attesting to their charges that decisions can be influenced by an appraiser's expectations about the appraisee's race, gender, or age rather than the appraisee's performance. After reviewing the litigation in performance appraisal, Cascio and Bernardin (1981) described their reactions to "reading the testimony of company officials in many of these cases. . . . One gets the uneasy feeling that (a) top management was totally unaware of what kind of appraisal system was in effect at lower levels, or (b) the employer was well aware of the (illegal) appraisal system in use but was unaware of what was wrong with it" (p. 223).

What Is Appraised

A major criticism concerns the quality of the information obtained about employees' performance and how these faulty appraisals

affect the quality of decisions made about them. Many of the conflicts center around the appropriateness of the *content* of the appraisal. Employees at the Internal Revenue Service, for example, griped about an overreliance on easy-to-measure production quotas over less readily quantifiable behaviors such as the judicious treatment of taxpayers. "In a numbers-driven organization, people find ways to meet the numbers. . . . If they say, 'You've got to collect X amount of dollars in a certain time,' well, all of a sudden, the taxpayer becomes subordinate to that goal" ("After Critical Inquiry," 1992, p. A18).

Some concerned parties go back a step to ask about the bases for deciding what is appraised: Who should decide? Can one rely on the expert opinion of subject matter professionals, or is it necessary to buttress the judgments of these esteemed experts with empirical evidence? If the latter, what type of validity evidence is required? In the case of *U.S.A. v. City of Chicago* (1978), for example, employees (designated as U.S.A.) were concerned about the evidence used in deciding whom to promote from Job A to Job B. Employees contended that it was a violation to use information from only Job A when there was little, if any, empirical indication of the overlap between Jobs A and B. In the case of *Brito v. Zia Co.* (1973), the court struck down the company's appraisal system because there were "no empirical data demonstrating that the appraisal system was significantly correlated with important elements of work behavior relevant to the jobs for which the appellants [employees] were being evaluated" (p. 1205). In ruling for the employees, the court actually specified the nature of the supporting or validity evidence.

The consequences of choosing an erroneous or irrelevant target can wreak havoc. In Kerr's (1975) aptly titled article, "On the Folly of Rewarding A, While Hoping for B," he presented case after case of administrators paying off for short-term earnings but hoping for long-term opportunities, and businesses dispensing rewards for unit performance but hoping for overall effectiveness.

How Performance Is Appraised

Another chronic complaint about appraisals concerns their *method,* that is, how the information is obtained: whether it is collected frequently enough, in what way, and by whom.

Regrettably, in most organizations, appraisals are done only annually or at best semiannually. For instance, one third of the employees in a manufacturing plant reported on an anonymous questionnaire that their performance was not evaluated at least once a year (Landy, Barnes, & Murphy, 1978). Among the reasons for the isolated appraisals is what is known as the "mum effect," that is, the tendency to keep mum about unpleasant messages. Tesser and Rosen (1975) detailed the ubiquity of this response in which people, in general, dislike transmitting negative information to a person who is directly affected by the message. In study after study, they show that the phenomenon occurs over a wide variety of "settings, communicators, recipients, messages, and indices of transmission" (p. 200). I/O psychologists report similar findings, particularly with subordinates who have performed poorly (Fisher, 1979; Ilgen & Knowlton, 1980; Larson, 1986). Bosses were found to minimize, or even distort, the negative aspects of their message.

Another concern centers on the appraisers themselves. Can managers be trusted to obtain accurate evaluations given their sometimes less-than-systematic methods of collecting information and potentially problematic biases? A candid set of interviews with 60 upper level executives vividly suggested that the unspeakable does happen: Truthfulness is not always their foremost concern

(Longenecker, Sims, & Gioia, 1987). Besides the slants introduced by such factors as gender, race, and age (Williams, 1997), many executives admitted that "political considerations *nearly always* were part of their evaluation process" (p. 183). They acknowledged giving "deflating" appraisals to shock an employee "to get him back on track" (p. 189) or to show followers "who the boss is" (p. 189). Some managers confessed to "inflating" the appraisals, not wanting to dampen the spirits of an employee who was belatedly improving or to avoid "hanging dirty laundry out in public" (p. 188). Executives may have a valid point when they argue that certain appraisals of employees may be blown out of proportion when written down or formalized. And political forces certainly exist in any organization. But employees rightfully contend that these defective appraisals affect the quality of decisions made about whom to promote, train, transfer, reward, or discipline.

WHY DO THE CRITICISMS PERSIST?

Widespread agreement exists that the quality of performance assessment is critical in both the ABA (Bellack & Hersen, 1988; Ciminero, Calhoun, & Adams, 1986; Goldfried & Kent, 1972; Johnston & Pennypacker, 1993) and I/O communities (Campbell, 1990; Cardy & Dobbins, 1994; DeNisi, 1996; Dunnette, 1963; James, 1973; Landy & Farr, 1983; Latham & Wexley, 1994; Murphy & Cleveland, 1995; D. Shaw, Schneier, Beatty, & Baird, 1995). Substantial space has been devoted to the topics of interrater reliability (1977, Vol. 10) and social validity (1991, Vol. 24) in the *Journal of Applied Behavior Analysis.* Behavior-analytic researchers recognize that the first two steps in any intervention—the specification and measurement of targets—profoundly influence the later step of modifying the behavior (Kent & Foster, 1977). In I/O psychology, performance appraisal is one of five major areas, and the "criterion problem" is singled out as a central issue. Moreover, Blum and Naylor (1968) proclaim the criterion as "basic to all measurement in industrial psychology. To overstate its importance would be literally impossible" (p. 174).

Expediency Rears Its Ugly Head

So why do the criticisms persist? One answer is the lure of expediency and tradition. An I/O psychologist with experience in industry and academia laments that we still "tend to take whatever criteria are available or are approved by management" (Thayer, 1992, p. 103). Furthermore, a wealth of sophisticated indices exist—sales volume, the number of lost-time accidents per million hours worked, stock-price performance—with few if any standards for how to define them. So strong is the pull of these indices that the "Uniform Guidelines on Employee Selection Procedures" (1978) caution against using an index simply because it is available.

Advice Given Falls Short

Few guidelines. A second reason is that few constructive alternatives and even fewer guidelines exist (Foster & Cone, 1986). The operational definition process, although acknowledged as critical (Bridgman, 1927), is seldom reported. Bormuth (1970) describes the process of selecting and creating items as "intuitive" (p. 56). Underwood (1957) laments in his classic research methods text that he "knows of no source to which one can turn which actually evaluates detailed matters involved in constructing an operational definition" (p. 51). This is still, unfortunately, true today. The reports by McClelland, Atkinson, Clark, and Lowell (1953), describing how they developed the Need for Achievement scale, and Pritchard, Jones, Roth, Stuebing, and Ekeberg (1988), detailing how they developed the ProMES

measurement system in the U.S. Air Force, are memorable in part because of their rarity.

Lack of specificity. Suggestions, although well meaning, are often found wanting. Much of the advice is vague. Let us take an example from the I/O psychology literature. In attempting to reconcile the overlapping and intertwined terms of construct-, content-, and criterion-related validity, Binning and Barrett (1989) suggested a unifying framework, including the relationship between the criterion measure, the performance domain, and the underlying psychological construct domain. In the performance domain, Binning and Barrett distinguished between outcomes "valued by the organization" and behaviors that are "the means to these valued ends" (p. 486), and they emphasized "the relevance of this distinction for criterion development" (p. 485). They went further to identify the procedures that personnel professionals traditionally use in deciding which behaviors and outcomes to include or exclude in the criterion measures for the job analysis. But they criticized the "idiosyncratic" use of job analyses and the "lack of general principles to guide data collection" (pp. 485–486). Binning and Barrett concluded by noting, "Perhaps the greatest advancement for the science of personnel psychology will come only when the values driving organizational administrators' decisions about behavioral science research are changed" (p. 490). Exactly what these changes should be and how they will affect what we do, however, remain unclear.

Reliance on evaluation sources. Another problem is that the advice given sometimes goes no further than specifying the sources that evaluate the targets. A case in point is the newest validity—social validity—promoted by behavior-analytic researchers (Bem & Funder, 1978; Cone, 1980; Schwartz & Baer, 1991; Winett, Moore, & Anderson, 1991; Wolf, 1978). They encourage their colleagues to see whether the implemented targets (as well as the treatment and outcomes) have a positive and meaningful impact on the client. To do so, researchers recommend actively seeking external guidance to buttress the opinions of professionals. The two suggested sources are (a) clients or consumers (e.g., mine workers, owners, and the families and friends of the direct consumers, as well as the insurance company and taxpayers and merchants in the community; Schwartz & Baer, 1991; Wolf, 1978) and (b) data collected about the target. For the latter, for example, Green and Reid (1996) gathered ratings from practitioners about their singular target of "happiness" and assessed whether they coincided with the observed indices of their clients. Researchers have also obtained epidemiological data, connecting a target such as physical activity with a risk such as heart disease (Winett et al., 1991).

Yet, sources disagree with one another. Consumers do not always speak in one voice, and they sometimes differ with professionals. A vivid dispute is portrayed in Hawkins' (1991) article on social validity in which a teacher recommended that a young mentally retarded man learn to brush his teeth. His parents disagreed, saying it was not worth the effort, and had all his teeth pulled out. Moreover, introducing the *JABA* issue on social validity, Geller (1991) contends that there are times when a consumer's preference "should be disregarded for societal benefit" (p. 180). Schreibman (cited in Geller, 1991) posed the provocative question of what we should do when

> the consumer is wrong? For example, if we were to acquire data from parents on the acceptability of a school program for developmentally disabled children, we might find that the most acceptable program is one that focuses on teaching the children to be cooperative and compliant to demands. . . . Given that we (as professionals) know im-

> provement in language is a significant prognostic indicator while compliance is not, what do we do? (p. 182)

Unfortunately, the answer to her question is not clear.

A NEW SET OF CRITERIA FOR CRITERIA: SURF & C

In response to the persistent criticisms about the quality of existing targets, I propose five criteria, referred to (and coined by Brendan O'Flaherty) by the mnemonic SURF & C:

S: Are the targets directly *sampled* rather than relying on filtered or secondary sources?

U: Are the targets primarily *under* the control of workers (or any persons being targeted), responsive to their efforts, and minimally affected by extraneous factors?

R: Do independent observers consistently agree on their recordings and obtain interrater *reliability* scores of 80% to (ideally) 90% or better during the formal data collection period?

F: Are the targets assessed *frequently*—at least 20 and ideally 30 times—during the period of the intervention?

C: Is there evidence indicating that the targets are *critical* for the successful completion of the task?

When the answer is yes to each of these five questions, then the target is considered to meet the SURF & C criteria.

Four of the criteria (SURF) are drawn from earlier work (Komaki, Collins, & Temlock, 1987); one (C) is discussed here for the first time. Many of the examples presented here take place in work settings, but the standards can be used in virtually any applied setting.

A Case

To illustrate how and why to use the SURF & C criteria, I describe two year-long field experiments. (Year 1 is reported in Komaki & Collins, 1982. Year 2 is detailed here for the first time.) Because this depiction focuses on my role, I give a first person account.

The impetus for the studies came from Marine Corps officers who were concerned about the preventive maintenance (PM) of equipment involving the adjusting, oiling, and replacement of suspiciously worn parts. Costly equipment breakdowns and replacements were occurring. Previous attempts, using a quality circle approach, had not been successful. I proposed using the same three-step approach I had used in promoting safety (Komaki, Barwick, & Scott, 1978). A contract was let through the Office of Naval Research; the site chosen was the Ordnance and Motor Transport sections of a battery in a heavy artillery battalion in a division at a Marine Corps base camp. In the battery, there were approximately 115 first and second echelon personnel.

The major challenge was the measurement of PM. Although PM had been cited again and again as a major factor (in this case) in airline accidents (National Transportation Safety Board, 1994), it had been difficult to motivate even the most highly trained work force to do it (Higgins, 1988). Demonstrations of actual on-the-job improvements are rare (Maggard & Rhyne, 1992; Ola d'Aulaire & Ola d'Aulaire, 1986; Wilkinson, 1968) in part because of the formidable measurement problems.

First, PM does not lend itself to traditional outcome assessments. It is not only intangible, but also has few immediate outcomes. Failing to inspect a jeep, for example, does not directly affect its operation. The jeep essentially looks and operates the same. Because supervisors cannot readily tell if PM has been done properly, short of trailing each employee constantly, they often end up relying on paperwork supposedly listing "repairs that needed to be done a second time"

(Bryant, 1995, p. A16). Unfortunately, this paperwork is one step removed from the work itself, and tricking the inspectors is not that difficult. " 'We were really good at . . . making the paperwork look correct,' one former airline employee testified at a Senate subcommittee, although what really happened on the airplane might be totally different" (Wald, 1996, p. E22). Some managers have been accused of falsifying PM records (Cushman, 1992; Salpukas, 1991); criminal charges were lodged against the managers of Eastern Airlines for falsifying maintenance logs (Weiner, 1990, p. 21).

Another contributing factor is the complexity of PM (March & Simon, 1958; McCann & Ferry, 1979; M. Shaw, 1973; Thompson, 1967; Victor & Blackburn, 1987). First echelon personnel in the Marine Corps start the PM chain and inspect the equipment for deficiencies. Second echelon personnel stock, order, and replace parts. Coordination failures are common. For example, the driver may find the temperature gauge inoperative and record its status on the log, but then fail to turn in the paperwork to the order clerk. Or the clerk may obtain the gauge, but then fail to forward the gauge to the mechanic. Hence, it is often tricky to determine why, after a month, the gauge is still inoperative.

Year 1: Developing targets. To obtain ideas for measuring PM, my colleagues and I searched the literature (Komaki & Penn, 1982) and talked with on-site personnel. Finding the traditional PM indicators (as shown in Table 1) unsatisfactory, we designed three new ones: (a) time utilization, (b) supervision, and (c) follow-through. We set up a program that included graduated goals and a potent consequence: time off with pay, shown in a previous military setting, in an article by Datel and Legters (1971), to be among the top-rated consequences. Surprisingly, improvements were either short-lived or nonexistent (Komaki & Collins, 1982).

Year 2: Reevaluating targets. The failure of Year 1 (and the not-unrealistic fear that my contract would be terminated) propelled me to do a reexamination. Among the things I discovered was that troops were sometimes ordered to be elsewhere when we collected data (during scheduled PM times), causing a problem with responsiveness.

The next year I was able to set up another experiment in another battery. To avoid problems with responsiveness, I momentarily considered the completion of PM logs. The Marines could fill out the logs any time, and they would therefore be under (U) their control. The Marines quickly impugned the accuracy of these "paper PMs," however, reminding me of what the Marines referred to as "pencil-whipping." More than one Marine admitted to filling out a PM log without so much as lifting the hood of the jeep. What was important, they emphasized, was not merely reports of deficiencies on the logs, but accurate reports. Because of the difficulties in establishing the standard (or "true score") against which to assess accuracy, I dreaded developing an accuracy measure. Given the importance of accurately identifying deficiencies, however, I reluctantly decided on two targets (identified in Table 1): (a) detected deficiencies: reflecting whether Marines correctly reported deficiencies, and (b) follow-through: identifying the action taken on the detected deficiencies. Only after considerable field testing and revamping of the PM logs (as shown in Table 2) was reliability obtained. Finally, each of the criteria were met:

S: Trained observers went on-site and directly sampled (S) the outcomes of the Marines' work (e.g., the deficiencies reported, the deficiencies verified, the repairs made, the parts ordered).

U: The targets were designed to be relatively unaffected by extraneous factors. With

Table 1
Preventive Maintenance (PM) Targets

Traditional	S	U	R	F	C	Year 1	S	U	R	F	C
Deadline rate Percentage of combat-essential equipment judged to be inoperative by supervisory personnel.	+	−	−	?	?	*Time utilization* Mean number of persons on task each minute, where on task referred to manipulating equipment with hands or tools.	+	−[a]	+	+	−
Maintenance cost to total operating cost The expenditure in dollars to maintain equipment relative to the expenditure to operate it, an index primarily used in the private sector.	NA	−	−	?	?	*Supervision* Percentage of time a supervisor was present, specifically within 10 m of the vehicle or gun being worked on whenever troops were present.	+	−[a]	+	+	?
Completion of paperwork/PM logs Completion of forms in which personnel rate equipment parts as satisfactory or deficient.	?	−	−	+	−	*Follow-Through 1* Percentage of items needing attention that were either corrected or for which the necessary paperwork had been processed.	+	+	+	+	+
Shop appearance Judgments about tidiness of work areas.	+	+	−	?	−						
Written knowledge tests Responses to multiple choice questions typically given during PM training.	−	+	+	−	?						

Note. + = yes, − = no, ? = sometimes, depending on the situation, NA = not applicable.
[a] Initially thought to be responsive because workers could be on task whenever they were present during scheduled PM periods. This assessment was later changed because workers were sometimes ordered to be elsewhere during PM periods.
[b] *Verified* refers to the number of deficiencies the trained mechanics found when inspecting the equipment. *Reported* refers to the number of deficiencies that first echelon report on the weekly PM Checklists and are in agreement with the verified number of deficiencies.
[c] The denominator for follow-through is the same as the numerator for detected deficiencies.

detected deficiencies, for example, a truck could have 3 or 30 deficiencies. Yet each Marine could obtain a score of 100% if he accurately identified all the deficiencies. No one was responsible for the deficiencies per se, only for reporting them. Hence, the targets were considered under (U) the Marines' control.

R: Reliability was assessed 16% to 18% of the time during the formal data collection. For detected deficiencies, the agreement scores ranged from 87% to 90% and, for follow-through, there was 100% agreement.

F: Data were collected frequently (F): approximately every other week for a total of 21 to 23 times during a 35-week intervention period in one group and for 15 to 16 times during a 25-week intervention period in the other group.

C: Data collected by following the PM chain at the beginning of the year showed that unless deficiencies were accurately detected, it was impossible to follow through and rectify them. Over the year, I also discovered that as more deficiencies were detected and more follow-throughs were successful, the actual deficiencies, one of the ultimate PM goals, declined over time (as confirmed by the results of an autoregressive

Table 1
(Extended)

Year 2	S	U	R	F	C
Detected deficiencies Percentage of equipment deficiencies correctly reported by first echelon personnel. It is calculated as:[b] $\left(\frac{\text{reported \# of deficiencies}}{\text{verified \# of deficiencies}}\right) \times 100$	+	+	+	+	+
Follow-Through 2 Percentage of deficiencies for which appropriate and timely action was taken. The actions could be any of the following: –repairing or adjusting the item –ordering the replacement part, or –processing the paperwork for repair. If any one of these actions is taken within a week, follow-through was considered successful. Follow-through is calculated as: $\left(\frac{\text{\# of deficiencies for which action was taken}}{\text{reported \# of deficiencies}^{c}}\right) \times 100.$	+	+	+	+	+

analysis). These two pieces of evidence lent credence to the targets being critical (C).

Significant improvements were found, even though only feedback was provided. A multiple baseline design across groups was used. In the Ordnance group, both detected deficiencies and follow-through significantly improved when and only when the intervention was introduced, from an average of 26% to 51% and 17% to 76%, respectively. In Motor Transport, follow-though significantly improved from an average of 23% to 54%. For detected deficiencies, there was a 6-week delay in improvements; discussions with on-site personnel after the 3rd and 4th weeks revealed a lack of proper supervision. (This delay instigated my now almost two-decade-long focus on supervisory personnel, described in Komaki, 1998.) When the supervisors implemented the originally recommended procedures, however, performance significantly improved from 25% to 60%. (The results of an autoregressive analysis, based on Gottman's, 1981, linear model, essentially confirmed the interpretations based on a visual inspection of the data.)

The reactions of on-site personnel were positive. Supervisory personnel rated the intervention as *very* to *extremely* effective and recommended that a similar system be instituted throughout the Marine Corps. All parties agreed that they had a better idea of the maintenance effort. One unit supervisor remarked that the targets were "probably as objective as any evaluation could be."

Lessons learned. In retrospect, I can see how I had fallen prey to expedience again and again. First, I was drawn to the utilization of time and the completion of PM logs because they were easily and reliably measured. The former was actually defined in four words: "manipulating tools or equipment"; it entailed no specialized knowledge, and after a single day the observers obtained interrater reliability. Second, I delayed using an accuracy measure during Year 1. I was not unaware of its importance, but I knew that I would have to spend 10-fold the time in order to ensure reliability. As it was, I ended up spending 2 months negotiating the subtle but essential additions to 200-plus definitions of items such as brake fluid levels (Table 2). Given how tainted my initial choices of targets had been, I vowed that I would attempt to articulate criteria that would help to counterbalance these ever-ready temptations.

Criteria Indicating What to Measure

To see how the criteria can be used, let us compare targets that differ in relation to how well they meet the criteria.

Critical (C). Many targets are chosen arbitrarily without any evidence that the target is important to the task at hand. This was definitely the case with time utilization. Un-

Table 2
Year 2: Detected Deficiencies Data Sheet and Revamped PM Log

Instructions: Check whether each item is satisfactory or deficient.[a]		
Equipment	Equipment items[a]	Activities (and criteria[b]) for deficiencies
Howitzer	Mounts	A. Check night lighting devices for operation and broken wires and knobs. B. Check all mounts and counters for proper operation. C. Check for moisture and mold in counter windows. D. Check leveling bubble vials to ensure bubble moves, vials are not cracked, and vial covers are present and movable.
Jeep	Clutch and brake pedals	A. Ensure pedals operate normally. B. Check clutch for excessive play (1.5 in. maximum). C. Check brake for excessive play (0.25 in. maximum).
Goat	Brake fluid	A. Open master cylinder to check fluid level (can see or touch with tip of finger, but not filled to top).
Truck	Air tank leaks	A. Run engine until 90 lb. pressure, wait 30 seconds, press brake pedal. Check for excessive play (4 in. maximum) or hissing.
Trailer	Tire pressure	A. Check for proper pressure (25 lb. for jeep trailer, 35 lb. for water and truck trailers).

[a] On original PM log (checklist).
[b] Added as a result of repeated tests of interrater reliability and discussions of disagreements.

fortunately, little if any evidence indicated that spending time this way would contribute in a meaningful way to PM. The same can be said for other often-used PM targets (shown in Table 1), such as the completion of paperwork, the appearance of the shop, and the results of written knowledge tests.

To obtain evidence, I recommend a contrasted or extreme groups design (Barron, 1955; Cowan, Conger, & Conger, 1989; Gilbert, 1978; Meyers, 1972) in which a group known to possess a certain characteristic is contrasted with a group lacking it. To generate safety targets, we watched neophyte and seasoned employees as each performed the same operation (Komaki, Collins, & Penn, 1982). The targets were the specific, detectable differences between the groups—in this case, the timing and motion of their actions. Others use this same strategy to identify desirable verbal social skills (Holmes, Hansen, & St. Lawrence, 1984) and actions that reduce the likelihood of heart attacks (Bloom, 1988).

Not just any evidence will do, however, as recent court cases have attested (*Brito v. Zia Co.,* 1973; *U.S.A. v. City of Chicago,* 1978). The only acceptable evidence is data showing that a meaningful relationship exists between the target and the ultimate goal (or in the parlance of an I/O psychologist, between the operational definition and the ultimate criterion). Reber and Wallin (1983) provided an excellent example when they validated an observational measure of safety. These researchers collected data for over 3 years on their target and their ultimate criterion of reducing accidents and found that the target correlated significantly and negatively with accidents. Other investigators, using the same target, could collect commensurate information or cite this article as evidence. Citations about the importance of safety in general, however, would not be sufficient to show that the target is critical. Opinions are also not sufficient, no matter who voices them—the CEO of the company or your boss—unless he or she bolsters them with data about the specific target. In this way, the requirement to provide validity ev-

idence counteracts the all-too-frequent expediency that creeps in when choosing targets.

Under control of workers (U). Another problem with many targets is their lack of responsiveness to workers' efforts. A typical way of judging PM (as shown in Table 1) is to look at the status of the equipment (deadline rate) or the costs associated with maintaining the equipment, and then to infer that PM is faulty if the equipment is down or the costs high. None of these indices would meet the criteria of being under workers' control because they are influenced by extraneous factors such as the age, use, and design of the equipment, the supply and procurement system, and the state of the economy.

Is it generally the case that outcomes are less responsive and that behaviors guarantee that targets will be under the control of workers? Not necessarily. Many targets commonly used in industry are outcomes: sales volume, stock prices, and plant productivity. Because they are often influenced by extraneous factors (e.g., merchandise mix, economic conditions, bad parts), they are usually not recommended as targets (Baker, Gibbons, & Murphy, 1992; Smith, 1976; Zipser, 1996). Not all outcomes are unresponsive: Detected deficiencies and follow-through are outcomes. With the latter, I ensured that the target would mirror the workers' efforts when I defined it as one of three outcomes, two of which were under their control (Table 1). The important distinction is not whether the target is a behavior or an outcome, but whether the target sensitively reflects a systematic and sustained amount of effort.

Criteria Concerned with How to Measure

Directly sampled (S). Many targets rely on indirect rather than direct assessments. Supervisors might rely on secondhand information rather than surveying the situation themselves; for example, instead of seeing firsthand whether taxpayers are treated fairly, one might rely on taxpayers making a formal complaint. Unfortunately, these complaints are at least one step removed, because even though taxpayers do not file a formal complaint, it does not mean that they were treated equitably. Because direct sampling of the actual behaviors or outcomes has been shown to result in more accurate and unfiltered information (Bernard, Killworth, & Sailer, 1979–1980; Burns, 1954; Hammer, 1985; Lewis & Dahl, 1976), we recommend using these indirect indices only as supplements.

Reliable (R). Whether it be the ubiquitous supervisory ratings or written knowledge tests, the assessment of interrater reliability in most work settings is regrettably rare. When reliability is assessed, agreement is often unacceptably low. For example, reliability checks of the limited technical inspections conducted at my request each week during Year 1 showed that two technically competent inspectors who had inspected the same piece of equipment within a 24-hr period disagreed 25% to 50% of the time.

To meet the reliability standard, reliability scores of at least 90% for an established measure and at least 80% for a new measure (Miller, 1997) must be obtained during the formal data collection on approximately 10% of the observations. Interrater reliability is typically calculated as a percentage figure: number of agreements divided by number of agreements plus disagreements and multiplied by 100% (for other ways of calculating it, refer to Foster & Cone, 1986).

To enhance reliability, I conducted tests of reliability during developmental and training stages. In order to be considered trained, for example, each mechanic had to obtain three consecutive, representative reliability scores of at least 90% or better before he was ready to collect data formally.

Conducting reliability checks during the developmental process provided natural springboards for ferreting out disagreements and their bases. These discussions should not be underestimated. With the Marines, for instance, disagreements often provoked questions such as, "How much play is too much play?" (for the status of brakes). Rather than defending one's ideas or disparaging another's, the focus was on making revisions so that the new definition captured, in this case, what constituted a deficiency, and we restated it in such a way ("1/4 in. maximum") so as to reduce future disagreements (as shown in Table 2).

Can this process be used to mitigate the inevitable biases of appraisers, or is it relegated to the arcane world of academia? In *Price Waterhouse v. Hopkins* (1989), a classic case of gender discrimination, substantial disagreements occurred among partners of an accounting firm because some differed about what persons of a certain gender should and should not do, and because of different weights some partners placed on internal matters such as staff relationships and external matters such as procuring major contracts for the firm. It is enticing to speculate what might occur if the partners would use the test of interrater reliability as the basis for developing a new appraisal system. Doing this would no doubt generate discussion about what constitutes performance worthy for promotion as well as how to identify various acceptable combinations that could lead to more uniform evaluations on the next round.

Interestingly, Baer, Wolf, and Risley (1968) long foreshadowed these concerns about observers' biases when they asked, "not merely, was *behavior* changed? but also, *whose* behavior?" (p. 93). Usually, the assumption is that the appraisee rather than the appraiser is changing. As Baer et al. (1968) point out, however, this is not always the case. Perhaps they would recommend that a similar approach be taken as a first step in dealing with bias, an insidious problem that has found few truly workable and effective solutions (Eichenwald, 1996; Williams, 1997).

Frequently assessed (F). Another common mistake is to appraise too infrequently. I rejected one noteworthy candidate (the biannual or annual evaluations) because they were done only once or twice a year. To use this evaluation, its frequency would have had to be increased by 20-fold (an impossible goal), or the period of evaluation would have had to be extended to 20 years (which no one was willing to do).

The recommendation to make assessments at least 20 (and ideally 30) times during the intervention period, and before the intervention is considered a failure, is based on three considerations. Theoretically, we know that one of the most straightforward ways to increase the potency of an intervention is to increase its frequency (Miller, 1997). A review of behavior-analytic interventions in work settings showed that at least half of the interventions lasted 8 weeks, and information was provided in the majority of cases weekly and as often as daily (Komaki, Coombs, & Schepman, 1991). Southard et al. (1992) found that targets in which feedback was given more frequently have been associated with better results than those with less frequent feedback. Second, we know that increasing the frequency enhances the representativeness of the information obtained (Cronbach, Gleser, Nanda, & Rajarathnam, 1972; Miller, 1997). An expert on time-series analyses (Gottman, 1981) notes that 175 observations are "not excessive" (p. 312). Although he admits that a number this high is not always necessary, he points out that "it is an extremely risky business" to be using as few as "five points before and five points after intervention" (pp. 58–59). Another statistician (R. Millsap, personal communication, April 10,

1997) stipulates that at least 20 to 30 data points are necessary to discern trends reliably. Furthermore, the frequency of assessment makes a difference to the persons being evaluated. Landy et al. (1978) identified three factors that were related to employees' attitudes about the fairness and accuracy of their evaluations: (a) how often the appraisals were done, (b) whether supervisors had pointed out goals that employees should strive toward to eliminate weaknesses, and (c) how well the supervisor actually knew the subordinate's level of performance and job duties. Given these considerations, it would behoove us to ensure that targets are measured both frequently and reliably.

SO WHAT?

Are SURF & C Criteria Really New?

The answer to this question depends on your perspective. If you are already meeting the criteria, have little difficulty identifying why some targets are better than others, and think most targets meet the criteria, then you may not see these criteria as noteworthy. If, however, you, like me, struggle when generating appropriate targets, have trouble justifying your choice, question other persons' justifications, find fault with many targets, and think they can be improved, then you may be more likely to see their value.

One advantage of any set of criteria is that they "strive for relevance to principle" (Baer et al., 1968, p. 96), enabling us to go beyond specifics. In discussing conceptual systems, Baer et al. pointed out: "To describe the exact sequence of color changes whereby a child is moved from a color discrimination to a form discrimination is good; to refer also to 'fading' and 'errorless discrimination' is better" (p. 96). The same holds for the proposed criteria. To describe how a target such as follow-through was operationally defined is good; to refer also to the criteria as being under control is better. Reference to a higher order lifts the discussion from idiosyncratic details about PM to what standards should be used in judging dependent variables.

But merely proposing standards is not unprecedented. A variety of criteria—quantifiability, controllability, relevance, freedom from bias, reliability, practicality, and discriminability—have been suggested by I/O (Bernardin & Beatty, 1984; DeVries, Morrison, Shullman, & Gerlach, 1980; Fleishman & Quaintance, 1984; Kane & Lawler, 1979; Thorndike, 1949) and ABA researchers (Daniels, 1989; Gilbert, 1978; Mash & Terdal, 1981; Sulzer-Azaroff & Fellner, 1984).

Furthermore, little disagreement exists for any of the standards: directly sampling targets (Ayllon & Azrin, 1968); ensuring that targets are under the control of workers (Bernardin & Beatty, 1984; Daniels, 1989; Gilbert, 1978); guaranteeing the reliability of observers (Foster & Cone, 1986; Liebert & Liebert, 1995; Miller, 1997; Schmitt & Klimoski, 1991; Selltiz, Wrightsman, & Cook, 1976; Stone, 1978); ensuring the frequency of data collection (Cronbach et al., 1972; Gottman, 1981; Miller, 1997); or obtaining evidence of criticality (sometimes referred to as criterion validity data, as identified in Ghiselli, Campbell, & Zedeck, 1981).

Noncompensatory nature. So is there anything distinctive about the SURF & C criteria? Yes: They work as a group. What this means is that all of the criteria must be met. No one criterion can make up for another. Hence, the high reliability scores of time utilization cannot compensate for a lack of criticalness and responsiveness. This noncompensatory characteristic also assumes that a target must be sound not only in method but also in content.

Tailored to performance motivation. Another way in which the SURF & C criteria are unusual is that their boundaries are explicitly

drawn to be motivational. In fact, they are tailored to maximize the effectiveness of frequent and positive reinforcement of targets that are tied to the ultimate goals of the organization. At the same time, they are also worker (rather than management) centered. Criteria such as practicality and inexpensiveness, which are oriented toward the target developer or management, are downplayed in favor of the person whose target is the focus: the worker, student, client. For descriptive or taxonomic aims, other standards such as being mutually exclusive and exhaustive (Fleishman & Quaintance, 1984) are more appropriate. Similarly, when selecting employees from a pool of candidates, discriminability and sensitivity (DeVries et al., 1980) should rightfully take precedence. When the goal is motivational, however, the SURF & C criteria are highly recommended.

Aren't We Preaching to the Choir?

Perhaps *JABA* readers do not need to hear about these standards, not because any major disagreement exists with them, but because this is not a serious problem. Having used the SURF & C criteria to critique my own work, I know I often fell short. In assessing friendliness, for instance, we arrogantly assumed that smiling was key (Komaki, Blood, & Holder, 1980). No empirical rationale was cited, nor was it sought. Yet, my business school students incisively asked, "Where are the validity data?" They even identified empirical data that included hundreds of critical incidents from customers of airlines, hotels, and restaurants, providing evidence from the customers' point of view of satisfactory service encounters (e.g., Bitner, Booms, & Tetreault, 1990). The examples ranged from ways in which personnel handled failures to responding over and above the call of duty to customers with special needs; never once was smiling mentioned. Based on data such as these, I would now change my original and rather naive definition of service.

JABA authors, however, may have fared better than I. Hence, a graduate student, Mahmut Bayazit, and I examined 3 years (1994–1996) of articles about subjects and settings we are familiar with: employees in work settings. In critiquing six targets using the SURF & C criteria, we did not find criticalness to be a major problem. Five of six met this criterion. Three obtained their own validity data, sometimes under the rubric of social validity, and two made relevant citations to empirically derived standards or pertinent data. The most prevalent problem was frequency. For half of the targets, data were collected no more than three to four times during the intervention.

Our assessment, though limited, was mixed. If we did not look at the frequency of the measurement, the choir sounded heavenly. But the SURF & C criteria were met in only 50% of the targets (Brothers, Krantz, & McClannahan, 1994; Johnson & Fawcett, 1994; Shore, Lerman, Smith, Iwata, & DeLeon, 1995). Although this may not be considered a serious breach, it indicates that the choir may be off-key and in need of some fine-tuning.

FUTURE DIRECTIONS

More reports about what goes on behind the scenes are encouraged. Wolf, Kirigin, Fixsen, Blase, and Braukmann's (1995) article, describing how their failure to assess the satisfaction of key consumer groups severely hampered their dissemination efforts, is a noteworthy step in this direction.

An intriguing question concerns the trade-offs among the steps in the behavior-analytic approach (Frederiksen, 1982): (a) specify, (b) measure, and (c) provide consequences for desired targets. Because we often do not have the luxury of maximizing all of the steps at once, I recommend empirically

determining where we should best place our scarce resources. My prediction is that a trade-off exists, but there are limits to how poorly targets can be measured. I predict that a potent consequence, such as time off, cannot compensate for a target that is poorly measured (meeting only the SRF criteria). On the other hand, I would expect that a higher quality measure (meeting the SURF & C criteria) would make up for a lower potency consequence such as feedback.

Another recommendation is to conduct a series of experiments documenting the impact of using the SURF & C criteria on the quality of targets and the developmental process. As instrument developers use the criteria, I would expect more discussion about expedience and the substance of the targets.

In conclusion, exposing the typically hidden operationalization process shed considerable light on the alluring temptations of precedence and expedience and the prodigious amount of work involved in successfully counteracting these pulls. Despite some naive claims to the contrary (Doran, 1997), picking target behaviors is "neither simple nor unerring" (Bechtoldt, 1959, p. 621). The proposed SURF & C criteria should help to generate ideas about how to be less arbitrary in our choices of targets and less ambiguous about our reasons for selecting these targets. They also address how and what information should be gathered, two of the major criticisms of performance appraisals. In doing this, a more stable foundation is provided for successfully motivating others. Moreover, striving to meet these criteria should enhance the quality of the information we obtain and the decisions we make, thus bringing us ever closer to the vision expressed by Baer et al. (1968) of "a better state of society" (p. 91).

REFERENCES

After critical inquiry. (1992, April 7). *The New York Times,* p. A18.

Austin, J. T., Villanova, P., Kane, J. S., & Bernardin, H. J. (1991). Construct validation of performance measures: Definitional issues, development, and evaluation of indicators. *Research in Personnel and Human Resources Management, 9,* 159–233.

Ayllon, T., & Azrin, N. H. (1968). *The token economy: A motivational system for therapy and rehabilitation.* New York: Appleton-Century-Crofts.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91–97.

Baker, G., Gibbons, R., & Murphy, K. (1992). *Subjective performance measures in optimal incentive contracts.* Paper presented at Columbia University Workshop in Applied Microeconomics, New York.

Barron, F. (1955). The disposition toward originality. *Journal of Abnormal and Social Psychology, 51,* 478–485.

Bechtoldt, H. P. (1959). Construct validity: A critique. *American Psychologist, 14,* 619–629.

Bellack, A. S., & Hersen, M. (Eds.). (1988). *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.

Bem, D. J., & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychology Review, 85,* 485–501.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1979–1980). Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Network, 2,* 191–218.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Bitner, M. J., Booms, B. H., & Tetreault, M. S. (1990). The service encounter: Diagnosing favorable and unfavorable incidents. *Journal of Marketing, 54,* 71–84.

Bloom, B. (1988). *Health psychology: A psychological perspective.* Englewood Cliffs, NJ: Prentice Hall.

Blum, M. L., & Naylor, J. C. (1968). *Industrial psychology* (rev. ed.). New York: Harper & Row.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists' Press.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Bridgman, P. W. (1927). *The logic of modern physics.* New York: Macmillan.

Brito v. Zia Co., 478 F.2d 1200 (10th Cir. 1973).

Brothers, K. J., Krantz, P. J., & McClannahan, L. E. (1994). Office paper recycling: A function of container proximity. *Journal of Applied Behavior Analysis, 27,* 153–160.

Bryant, A. (1995, October 15). Poor training and discipline at F.A.A. linked to 6 crashes. *The New York Times,* pp. A1, A16.

Burns, T. (1954). The directions of activity and communication in a departmental executive group. *Human Relations, 7,* 73–97.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives.* Florence, KY: South Western.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology, 34,* 211–225.

Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (Eds.). (1986). *Handbook of behavioral assessment.* New York: Wiley-Interscience.

Cone, J. D. (1980). *Template matching procedures for idiographic behavioral assessment.* Paper presented at the meeting of the Association for Advancement of Behavior Therapy, New York.

Cowan, G. S., Conger, J. C., & Conger, A. J. (1989). Social competency and social perceptivity in self and others. *Journal of Psychopathology and Behavior Assessment, 11,* 129–142.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajarathnam, N. (1972). *The dependability of behavioral measures.* New York: Wiley.

Cushman, J. H., Jr. (1992, November 26). Delta given 2nd-largest fine for lax maintenance. *The New York Times,* p. A17.

Daniels, A. C. (1989). *Performance management. Improving quality productivity through positive reinforcement.* Tucker, GA: Performance Management Publications.

Datel, W. E., & Legters, L. (1971). The psychology of the Army recruit. *Journal of Biological Psychology, 12(2),* 34–40.

DeNisi, A. S. (1996). *A cognitive approach to performance appraisal: A program of research.* London: Routledge.

DeVries, D. L., Morrison, A. M., Shullman, S. L., & Gerlach, M. L. (1980). *Performance appraisal on the line.* Greensboro, NC: Center for Creative Leadership.

Doran, K. J. (1997, October 10). No equal candidates. Letter to editor; Merit test is divisive in affirmative action case. *The New York Times,* p. A22.

Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology, 47,* 251–254.

Eichenwald, K. (1996, November 10). The two faces of Texaco. *The New York Times,* pp. E2, F10–11.

Finney, J. W. (1991). Selection of target behavior and interactions: A case study of necessary but insufficient choices. *Journal of Applied Behavior Analysis, 24,* 713–715.

Fisher, C. D. (1979). Transmission of positive and negative feedback to subordinates: A laboratory investigation. *Journal of Applied Psychology, 64,* 533–540.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance. The description of human tasks.* New York: Academic Press.

Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed., pp. 253–324). New York: Wiley-Interscience.

Frederiksen, L. W. (Ed.). (1982). *Handbook of organizational behavior management.* New York: Wiley.

Geller, E. S. (1991). Where's the validity in social validity? *Journal of Applied Behavior Analysis, 24,* 179–184.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Gilbert, T. F. (1978). *Human competence.* New York: McGraw-Hill.

Goldfried, M. R., & Kent, R. N. (1972). Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin, 77(6),* 409–420.

Gottman, J. M. (1981). *Time series analysis: A comprehensive introduction for social scientists.* Cambridge, England: Cambridge University.

Green, C. W., & Reid, D. H. (1996). Defining, validating, and increasing indices of happiness among people with disabilities. *Journal of Applied Behavior Analysis, 29,* 67–78.

Hammer, M. (1985). Implications of behavioral and cognitive reciprocity in social data. *Social Networks, 1,* 189–201.

Hawkins, R. P. (1991). Is social validity what we are interested in? Argument for a functional approach. *Journal of Applied Behavior Analysis, 24,* 205–213.

Higgins, L. R. (1988). *Maintenance engineering handbook.* New York: McGraw-Hill.

Holmes, M. R., Hansen, D. J., & St. Lawrence, J. S. (1984). Conversational skill training with after care patients in the community: Social validation and generalization. *Behavior Therapy, 15,* 84–100.

Ilgen, D. R., & Knowlton, W. A., Jr. (1980). Performance attributional effects on feedback from su-

periors. *Organizational Behavior and Human Performance, 25,* 441–456.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin, 80,* 75–83.

Johnson, M. D., & Fawcett, S. B. (1994). Courteous service: Its assessment and modification in a human service organization. *Journal of Applied Behavior Analysis, 27,* 145–152.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of human behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. M. Staw (Ed.), *Research in organizational behavior* (Vol. 1, pp. 425–478). Greenwich, CT: JAI Press.

Kent, R. N., & Foster, S. L. (1977). Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 279–328). New York: Wiley-Interscience.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal, 18,* 769–782.

Komaki, J. L. (1998). *Leadership from an operant perspective.* London: Routledge.

Komaki, J., Barwick, K. D., & Scott, L. R. (1978). A behavioral approach to occupational safety: Pinpointing and reinforcing safe performance in a food manufacturing plant. *Journal of Applied Psychology, 63,* 434–445.

Komaki, J., Blood, M. R., & Holder, D. (1980). Fostering friendliness in a fast-foods franchise. *Journal of Organizational Behavior Management, 2,* 151–164.

Komaki, J., & Collins, R. L. (1982). Motivation of preventive maintenance performance. In R. M. O'Brien, A. M. Dickinson, & M. Rosow (Eds.), *Industrial behavior modification: A learning-based approach to business management* (pp. 243–265). New York: Pergamon.

Komaki, J., Collins, R. L., & Penn, P. (1982). Role of performance antecedents and consequences in work motivation. *Journal of Applied Psychology, 67,* 334–340.

Komaki, J., Collins, R. L., & Temlock, S. (1987). An alternative performance measurement approach: Applied operant measurement in the service sector. *Applied Psychology: International Review, 36,* 71–89.

Komaki, J. L., Coombs, T., & Schepman, S. (1991). Motivational implications of reinforcement theory. In R. M. Steers & L. W. Porter (Eds.), *Motivation and work behavior* (5th ed., pp. 87–105). New York: McGraw-Hill.

Komaki, J. L., & Penn, P. (1982). Better business through behaviorism: Welcome to the real world of preventive maintenance. *Behavior Therapist, 5,* 159–163.

Landy, F. J., Barnes, J. L., & Murphy, K. R. (1978). Correlates of perceived fairness and accuracy of performance evaluation. *Journal of Applied Psychology, 63,* 751–754.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* New York: Academic Press.

Larson, J. R., Jr. (1986). Supervisors' performance feedback to subordinates: The impact of subordinate performance valence and outcome dependence. *Organizational Behavior and Human Decision Processes, 37,* 391–408.

Latham, G. P., & Wexley, K. N. (1994). *Increasing productivity through performance appraisal* (2nd ed.). Reading, MA: Addison-Wesley.

Lewis, D. R., & Dahl, T. (1976). Time management in higher education administration: A case study. *Higher Education, 5,* 49–66.

Liebert, R. M., & Liebert, L. L. (1995). *Science and behavior: An introduction to methods of research* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Longenecker, C. O., Sims, H. P., Jr., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *The Academy of Management Executive, 1,* 183–193.

Maggard, B. N., & Rhyne, D. M. (1992, fourth quarter). Total productive maintenance: A timely integration of production and maintenance. *Production and Inventory Management Journal,* pp. 6–10.

March, J. G., & Simon, H. (1958). *Organizations.* New York: Wiley.

Mash, E. J., & Terdal, L. G. (1981). Behavioral assessment of childhood disturbance. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (pp. 3–78). New York: Guilford.

McCann, J. E., & Ferry, D. L. (1979). An approach for assessing and managing inter-unit interdependence. *Academy of Management Review, 4,* 113–119.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive.* New York: Appleton-Century-Crofts.

Meyers, J. (1972). *Fundamentals of experimental design.* Boston: Allyn & Bacon.

Miller, L. K. (1997). *Principles of everyday behavior analysis* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Murphy, K. R., & Cleveland, J. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

National Transportation Safety Board. (1994). *National Transportation Safety Board annual report to Congress, 1993.* Washington, DC: U.S. Government Printing Office.

Ola d'Aulaire, E., & Ola d'Aulaire, P. (1986). Their mission: Make damn sure nothing comes unraveled in the air. *Smithsonian, 17,* 48–54.

Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).

Pritchard, R. D., Jones, S. D., Roth, P. L., Stuebing, K. K., & Ekeberg, S. E. (1988). Effects of group feedback, goal setting, and incentives on organizational productivity. *Journal of Applied Psychology, 73,* 337–358.

Reber, R. A., & Wallin, J. A. (1983). Validation of a behavioral measure of occupational safety. *Journal of Organizational Behavior Management, 5(2),* 69–77.

Salpukas, A. (1991, March 2). Eastern air admits to conspiracy. *The New York Times,* pp. 31, 42.

Schmitt, N. W., & Klimoski, R. J. (1991). *Research methods in human resources management.* Cincinnati: South-Western.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis, 24,* 189–204.

Selltiz, C., Wrightsman, L. S., & Cook, S. W. (Eds.). (1976). *Research methods in social relations* (3rd ed.). New York: Holt, Rinehart, and Winston.

Shaw, D. G., Schneier, C. E., Beatty, R. W., & Baird, L. S. (1995). *The performance measurement, management, and appraisal sourcebook.* Amherst, MA: Human Resource Development Press.

Shaw, M. E. (1973). Scaling group tasks: A method for dimensional analysis. *JSAS Catalog of Selected Documents in Psychology, 3,* 8.

Shore, B. A., Lerman, D. C., Smith, R. G., Iwata, B., & DeLeon, I. G. (1995). Direct assessment of quality of care in a geriatric nursing home. *Journal of Applied Behavior Analysis, 28,* 435–448.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–776). Chicago: Rand McNally.

Southard, D. R., Winett, R. A., Walbert-Rankin, J. L., Neubauer, T. E., Donckers-Roserveare, K., Burkett, P. A., Gould, R. A., Lombard, D., & Moore, J. F. (1992). Increasing the effectiveness of the national cholesterol education program: Dietary and behavioral interventions for clinical settings. *The Society of Behavioral Medicine, 14,* 21–30.

Stone, E. F. (1978). *Research methods in organizational behavior.* Santa Monica, CA: Goodyear Publishing Co.

Sulzer-Azaroff, B., & Fellner, D. (1984). Searching for performance targets in the behavioral analysis of occupational health and safety: An assessment strategy. *Journal of Organizational Behavior Management, 6,* 53–65.

Tesser, A., & Rosen, S. (1975). *The reluctance to transmit bad news (1).* Athens: University of Georgia, Department of Psychology.

Thayer, P. W. (1992). Construct validation: Do we understand our criteria? *Human Performance, 5,* 97–108.

Thompson, J. D. (1967). *Organizations in action.* New York: McGraw-Hill.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Underwood, B. J. (1957). *Psychological research.* New York: Appleton-Century-Crofts.

Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg., 38290–38315 (1978).

U.S.A. v. City of Chicago, 573 F.2d 416 (7th Cir. 1978).

Victor, B., & Blackburn, R. S. (1987). Interdependence: An alternative conceptualization. *Academy of Management Review, 12,* 486–497.

Wald, M. L. (1996, May 5). No-frills airline safety. *The New York Times,* p. E22.

Weiner, E. (1990, December 15). Eastern airlines data said to be still falsified. *The New York Times,* p. 21.

Weist, M. D., Ollendick, T. H., & Finney, J. W. (1991). Toward the empirical validation of treatment targets in children. *Clinical Psychology Review, 11,* 515–538.

Weitz, J. (1961). Criteria for criteria. *American Psychologist, 16,* 228–232.

Wilkinson, J. J. (1968, March–April). How to manage maintenance. *Harvard Business Review, 46,* pp. 100–111.

Williams, M. E. (Ed.). (1997). *Discrimination: Opposing viewpoints.* San Diego, CA: Greenhaven Press.

Winett, R. A., Moore, J. F., & Anderson, E. S. (1991). Extending the concept of social validity: Behavior analysis for disease prevention and health promotion. *Journal of Applied Behavior Analysis, 24,* 215–230.

Winett, R. A., & Winkler, R. C. (1972). Current behavior modification in the classroom: Be still, be quiet, be docile. *Journal of Applied Behavior Analysis, 5,* 499–504.

Wolf, M. M. (1978). Social validity: The case for subjective measurement, or how behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203–214.

Wolf, M. M., Kirigin, K. A., Fixsen, D. L., Blase, K. A., & Braukmann, C. J. (1995). The teaching-family model: A case study in data-based program development and refinement (and dragon wrestling). *Journal of Organizational Behavior Management, 15,* 11–68.

Zagat New York City restaurant survey: 1997. (1996). New York: Zagat Survey.

Zipser, A. (1996, August 26). Sunnier skies for TWA. *Barron's,* p. 10.

Originally received July 31, 1992
First acceptance June 23, 1994
Revision received September 24, 1996
Initial editorial decision January 6, 1997
Final acceptance December 18, 1997
Action Editor, Richard A. Winett